PCIe ® Over Fibre Optics: Challenges and Pitfalls

Derek Percival
Senior Field Applications Engineer
Avago Technologies

Disclaimer

Presentation Disclaimer: All opinions, judgments, recommendations, etc. that are presented herein are the opinions of the presenter of the material and do not necessarily reflect the opinions of the PCI-SIG®.

Agenda

- PCIe® and Optical Interfaces
- Optical Interface Options
- Challenges of Using Optics
  - ✓ Receiver Detect
  - ✓ Quiet and Idle Time
  - ✓ 8 GT/s Challenges
  - ✓ Resets and Clocks
- Tuning Performance
- Sideband Signals
- Customer Issues and Examples
- Boards and Optics
- Summary

PCIe and Optical Interfaces

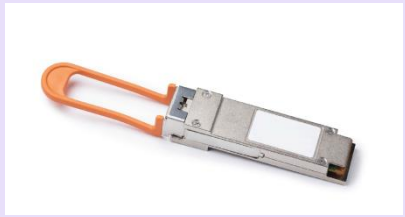
- Today there's no complete specification for PCIe over optical interfaces
  - ✓ Current optical interfaces are not optimised for PCIe use
  - ✓ Often designed for proprietary interfaces or other non-PCIe standards
  - ✓ This leads to issues with
    - Supported frequency ranges
    - Serdes levels and impedances
    - Squelching of quiet periods
    - Lane delays
    - Latency
    - Side band signals
  - ✓ Future cable specifications such as OCuLink and PCIe 3.0 Cable Spec include the possibility of using optical interfaces within them
    - As these are not yet released currently both specs are short on detail with regards to how AOC's will interface with PCIe devices
    - Methods listed here can potentially be used with these new standards

PCIe and Optical Interfaces

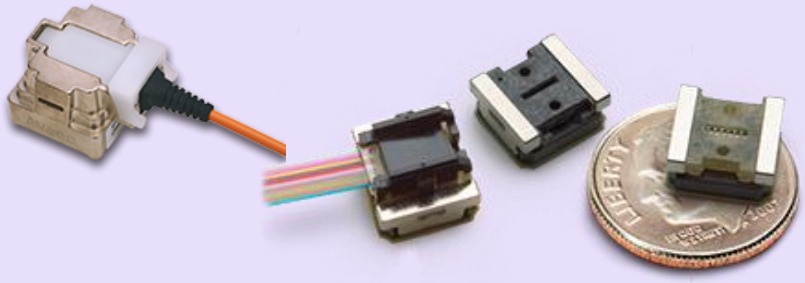
- Most testing has been done with Avago/PLX devices as link partners
  - ✓ Limited number of other PCIe devices have been shown to link
  - ✓ All optical links have required some “tuning” of the serdes to provide a reliable interface
  - ✓ Testing and evaluation is required on the optical modules used before rolling out a design
  - ✓ Reliable link up can be achieved but it requires some work
- Aim of this presentation is to provide you with hints and tips on how to achieve reliable links through optics

Optical Interface Options

- Many types of Optical interfaces can be used
  - ✓ We've tested many of these by many manufacturers
- Active Optic Cables tested include
  - ✓ SFP+
  - ✓ QSFP+
  - ✓ Mini-SAS HD
  - ✓ I-Pass
  - ✓ Etc.



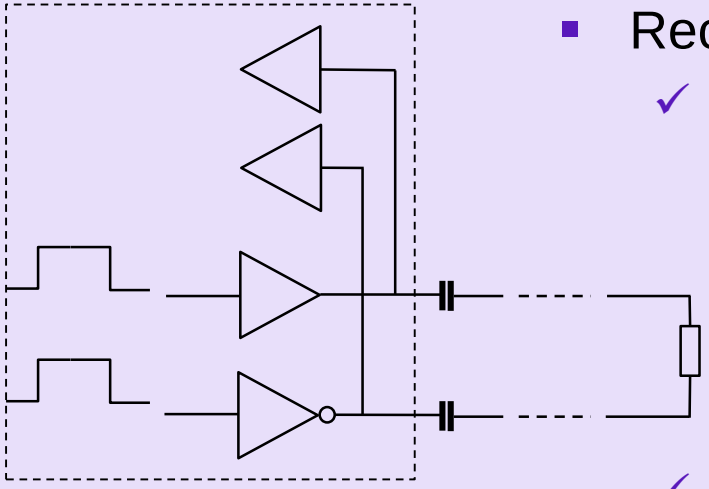
- Direct attached include
  - ✓ MiniPod
  - ✓ MicroPod
  - ✓ Etc.



Challenges of Using Optics

- Optical interface must be able to handle the frequency range of PCIe
  - ✓ Everything starts at 2.5 GT/s then switches to the appropriate speed
  - ✓ Optical interfaces have to handle not just the final link speed required but also the 2.5 GT/s link speed
    - Some optical interfaces require selection of the appropriate bandwidth range via straps
    - Important to ensure that 2.5 GT/s link speed is included within the same bandwidth range as the final target link speed
  - ✓ This is not normally an issue for most modules as they don't do re-timing

Receiver Detect



RD detection circuitry in TX detects change in pulse shape caused by AC coupling and termination

- Receiver Detect
  - ✓ Impedances of Optical interfaces often confuse RD circuitry
    - Generally terminated at 50 / 100 ohms, even unpowered in many cases!
    - Termination scheme can inhibit detection as may be presented in a non-compliant manner (cf PCIe spec.)
      - For example many are not terminated to Ground
  - ✓ If the Receiver Detect circuitry doesn't see the appropriate impedance
    - Will not begin link training – stays in Detect state
    - May link sometimes but not others
    - Some lanes may link and some may not reducing link width
  - ✓ Needs to be tested on the serdes/optics used in the system
    - If possible have the ability to mask receiver detect in the PCIe link partners

Quiet and Idle Time

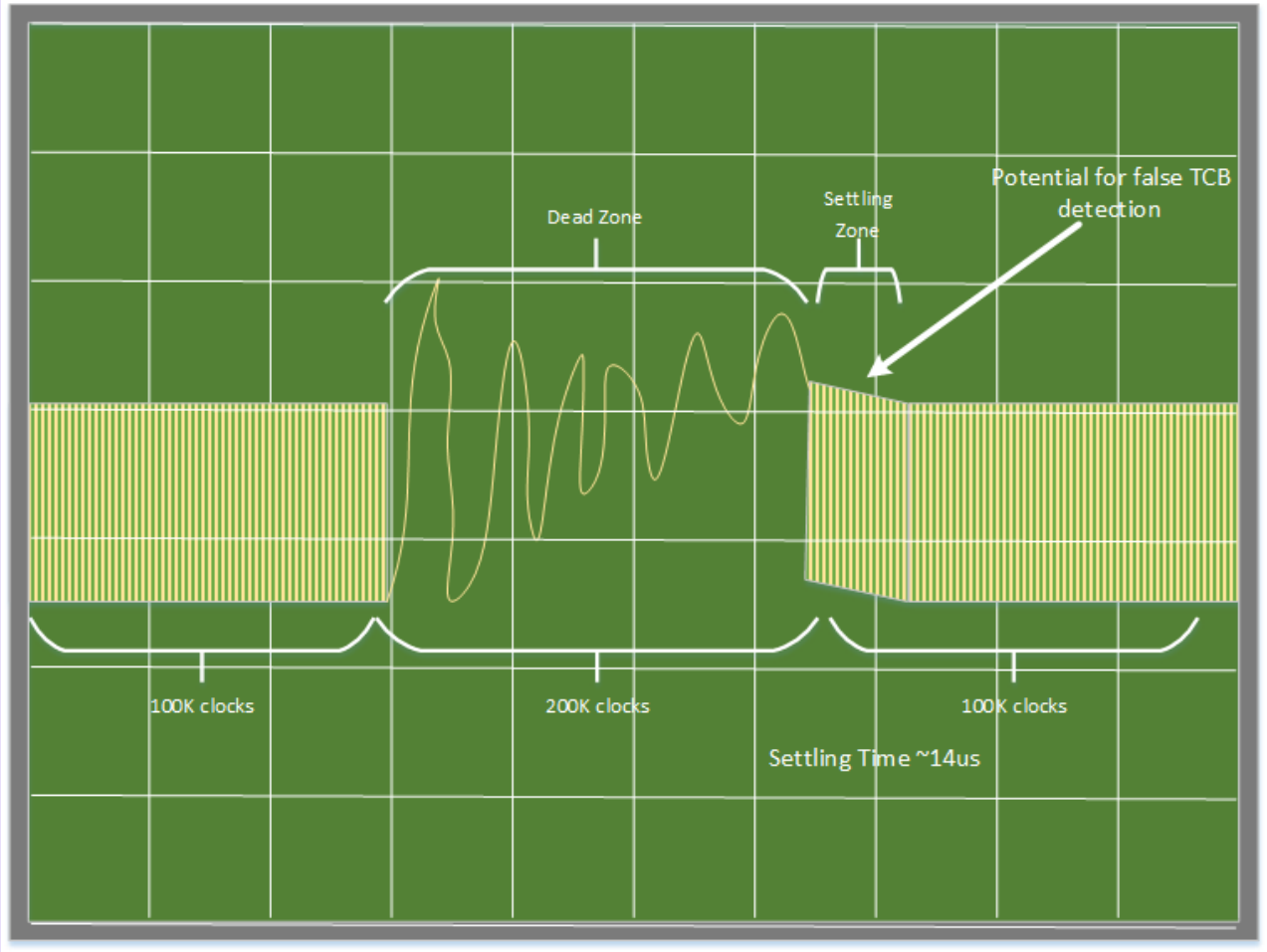
- Quiet or idle periods on the link where there is no traffic being sent can cause issues
  - ✓ This can occur when:
    - link is between detect and sending training sequences or
    - if the link goes into electrical idle caused by power management or
    - link is retraining due to speed or link width changes
  - ✓ Typical module specifies startup in milliseconds…
  - ✓ Result of this is:
    - Absence of optical modulation can saturate the PIN diode (RX)
    - Chatter can occur (No squelch)
    - Link can take significantly more time to stabilize due to
      - False locking
      - False symbol detection
    - Can result in false detection of Training Compliance Bits

Optical Saturation

- Example of ‘dead-time’ saturation in optics
  - ✓ Test example of Avago Minipod optics
  - ✓ Test consisted of a pattern generator feeding an Avago TX->Laser->PD->AvagoRX->Scope
  - ✓ Run patterns of varying non-transition and transition lengths to simulate varying times of EIDLE feeding the optics. Observe recovery time.
  - ✓ Test included multiple lanes and TX/RX on each Minipod
  - ✓ Several patterns tested but only one example shown

Optical Saturation

- Pattern is 100K clocks, 200K zero's, 100K clocks



Additional Saturation Comments

- Settling time can vary markedly between lanes on an optical link and between optical interfaces of the same model
- Early link noise can confuse PCIe serdes
  - ✓ Can delay link up time considerably
    - Maybe by minutes

Quiet and Idle Time cont'd

- Disable ASPM to prevent low power link states
  - ✓ Usually possible to set in device registers or in the BIOS/OS
- Disable autonomous link/width speed changes if possible
  - ✓ Usually possible to set in device registers
- Reducing the Detect.Quiet Wait Time if possible
  - ✓ Requires access to Phy settings
  - ✓ May not be needed depending on the Phy
- Disable TCB bits once compliance testing completed
  - ✓ Requires access to Phy settings
- Take care using Squelch or other optical serdes features
  - ✓ Well designed squelch and AGC should improve link up
  - ✓ Badly designed AGC may cause issues and add delays
    - – Delays can cause missed training sequences or malformed TLP's
  - ✓ Keep link active if possible to avoid need for these

8 GT/s Challenges

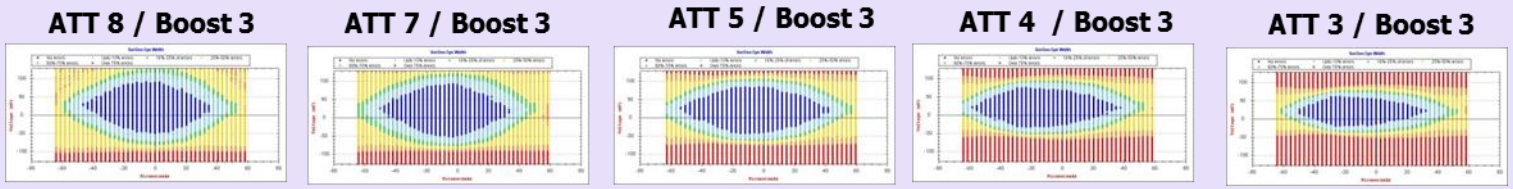
- Optical devices are non-linear
  - ✓ TX Side
    - TX backchannel tuning may not work or may give unreliable results
    - May need to bypass training phases
    - May need to use fixed presets or fixed coefficients
  - ✓ RX Side
    - Optical module and trace length dependent
    - It is likely that CTLE values of ATT & Boost will need to be either:
      - Modified to have a specific start value or
      - Locked to a fixed value

Tuning Performance

- Optical interface may require lower input levels to prevent receiver saturation
  - ✓ May need to reduce the overall transmitter amplitude levels
- If possible capture the eye on both sides of the link
  - ✓ Ideally use serdes built in eye capture after the equalisers
- Tune transmitter and receiver parameters to obtain best possible eye
  - ✓ In the case of 8 GT/s lock down transmitter and receiver parameters and prevent auto-tuning
  - ✓ May be able to get away with just setting start values of CTLE and allowing CTLE tuning in some cases.

Tuning Continued

- Bigger isn't necessarily better
  - ✓ Excessive amplitude in the receiver can cause saturation of the serdes slicer
  - ✓ Attenuate in the CTLE to ensure this doesn't happen
  - ✓ Eyes and their limits will vary between serdes vendors so contact serdes vendor for advice
  - ✓ Eye extraction is sub-sampling so check with error counters etc to determine link quality



Additional Changes

- In addition to disabling receiver detect, if possible use inferred electrical idle detection
  - ✓ Rely on data rather analog threshold to change speed
  - ✓ Forces TX training sets
  - ✓ Will require access to Phy registers
- Down-train Disable – inhibits width reduction
  - ✓ Will require access to Phy registers
    - Only needed on one end of the link
  - ✓ Optical lane ‘start up’ time varies
  - ✓ Downtrain disable forces retrain if needed
    - Link training may fail on some lanes if that optical lane isn’t ready
    - Retrain fixes this issue

Sideband Signals

- Optical interfaces may not have capability or spare capacity for:
  - ✓ PERST#
  - ✓ REFCLK
  - ✓ Power On
  - ✓ WAKE#

Sideband Signals

- Not obvious how you can push PERST# and other low frequency sideband signals down optics
  - ✓ Possible options are
    - Use Autonomous resets e.g. Power good on downstream devices/cards
      - Not ideal as cannot generate a PERST# to the remote device
    - Use a spare optical link(s) or run a sideband cable(s) along with the high speed optics
      - Increases cost of the interface

Sideband Signals cont'd

- Many optics have I2C/SMBus interfaces or sideband signals which can be used to assist
  - ✓ Use light presence and/or modulation present to generate PERST#
    - E.g. use GPIO or other signals to control optical enable at one end and use "light out" indication at the other end (perhaps AND'ing signals from lanes) to generate the PERST# or other signals
  - ✓ Use I2C/SMBus and special LED patterns to generate PERST#
    - May require FPGA or CPLD to provide additional logic
    - Can generate PRSNT# and other low frequency sideband signals in the same manner
  - ✓ These methods have been used/demonstrated by Avago/PLX and several customers
    - No standard yet but emerging standards do have access to the optics via I2C/SMBus e.g. Cable Management Interface in Gen 3 cable spec.
    - Currently locks user to one type of optics

Sideband Signals - REFCLK

- Same issue as for the other sidebands
  - ✓ i.e. may need another cable/optical link for the clock
- Asynchronous REFCLK operation possible
  - ✓ Pre-SRIS compliant
    - Need a low noise non-SSC REFCLK on each end of the link
    - SRNS compliant REFCLK should work but would need testing on pre SRIS devices
    - If you can't guarantee the noise or the REFCLK type use SSC isolation capable devices and run the REFCLK seperately
  - ✓ SRIS Compliant
    - Needs SRIS compliant REFCLK at both ends
    - Needs SRIS compliant devices at both ends
  - ✓ As long as REFCLK noise is low, link up is normally possible
    - Avago/PLX Switch to switch is very reliable
    - Asynchronous operation is the most commonly used method

Reliability

- Once tuned optical links have proven very reliable
- High reliability designs should ideally have:
  - ✓ an (optical) PERST# mechanism
  - ✓ and/or watchdog for link status
- Monitor Error registers to validate optimal settings and to check for changes in performance
  - ✓ Error counters are very useful for this and can be used to set alarms or bring the link down in the case of excessive errors

How Far Can You Go

- Longer optical cables add delay
  - ✓ Adds latency
  - ✓ Eventually affects ACK/NAK and updateFC protocol
    - Reduces performance
    - This affect also seen when adding re-timers or other passive delay elements
- Tested up to 50m of cable
  - ✓ 100m should be possible without issues
  - ✓ Few km (1-2) probably Ack/Nak limit but haven't tested
    - May be able to get further by bending the spec.
    - Likely to be throughput hits on an x4 link as you get over a 200m

Can You Use Any PCIe Device

- Most testing has been with Avago/PLX switches
  - ✓ All types over all generations
  - ✓ Serdes are highly configurable
- Off the shelf end points are unlikely to work without assistance from the vendor
  - ✓ Not seen any work yet
- Re-timers have been seen to work in one case but required lots of assistance from the vendor
- FPGA's have been seen to work but customer had to adapt the Phy settings as per the previous notes
- Custom solutions have been created to allow PCIe interconnect but require additional ASIC's or adapter cards

Customer Issues

- The link was working fine without masking receiver detection. Now it doesn't. Why do I now need to mask RD?
  - ✓ For some systems, we see TS1s driven from the PCIe device to the adapter and then after 24ms or so, it times out and sends a compliance pattern
  - ✓ When the adapter is replaced by another card or optical cable we go immediately into training without any compliance pattern sent

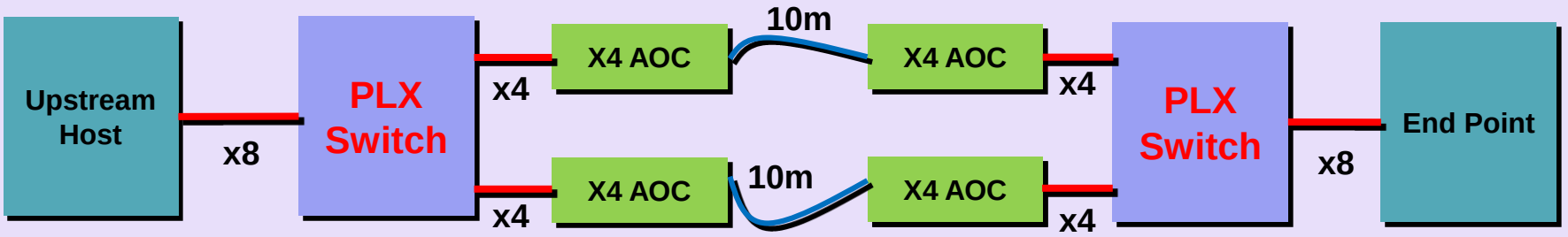
- Conclusion: Optical Modules are not intended for PCIe and there can be variability between modules (even within same model/manufacturer)
  - ✓ Masking RD allows for consistent link training across a range of devices/modules

Customer Example

- Using dual x4 AOC devices to create an x8 link
- Initial issue was inability to attain an x8 link
  - ✓ Achieved operation without bit error
  - ✓ Minimized link training time
  - ✓ Steps to achieve this were as per previous foils:
    - Basic masks – RX detection / EIDLE (Attain Link)
    - Down-train inhibit (Force max width)
    - TX/RX calibration (Link bit error reduction)
    - Detect Quiet delay reduction (improve multi-lane linkup time)

Link Fail Using Two Active Optical Cables

- One x4 AOC okay
- x8 using two AOC's fails to link correctly
  - ✓ Dual x4 Optical modules used to create an x8 Optical link
  - ✓ Independent Reset / Power testing
  - ✓ Max x8 Link time seen to be seconds to minutes for all 8 lanes!

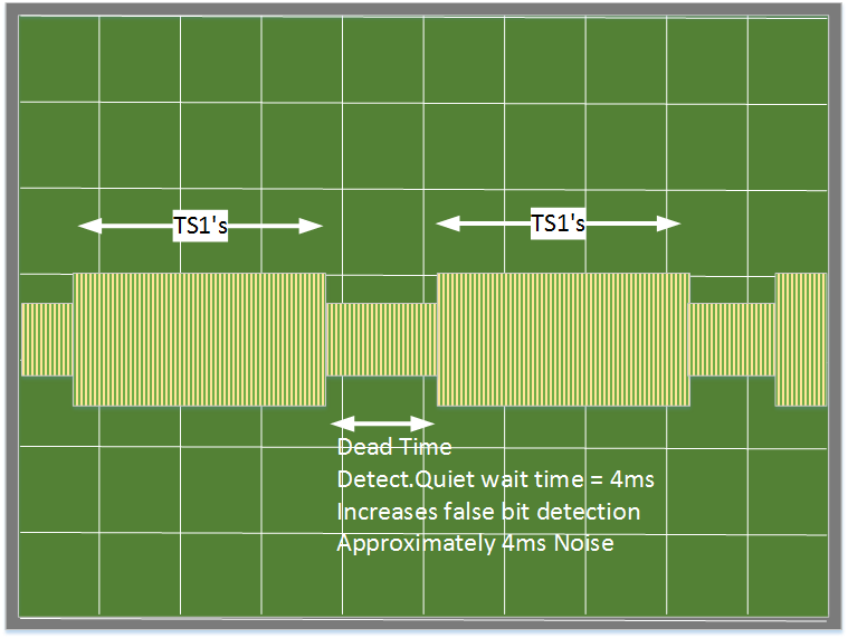


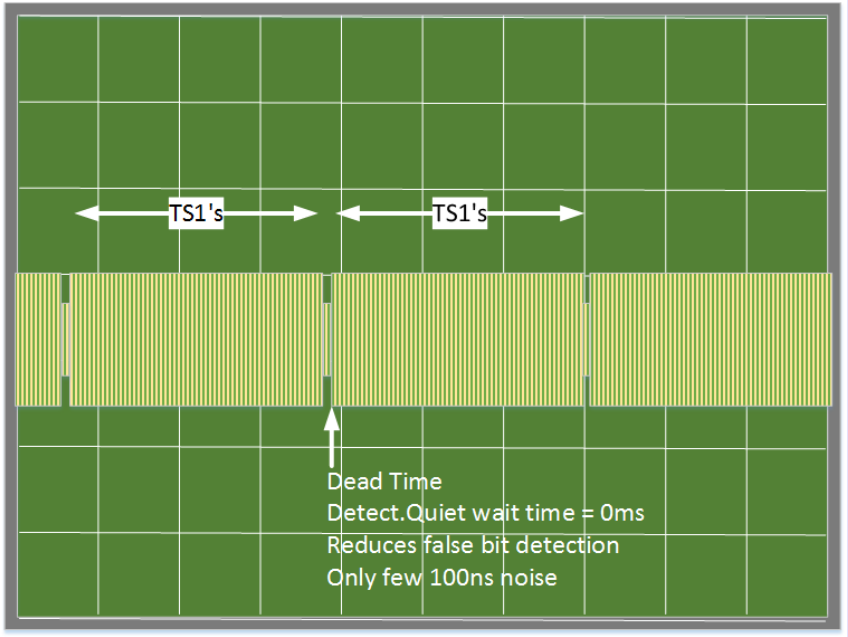
Steps To Get Link Up

- First test the optical modules to ensure these operate correctly
  - ✓ Test each AOC individually to
- As per previous foils
    - Switch to inferred Electrical Idle
    - Mask Receiver Detect
    - Disable down train
  - ✓ Link now comes up as x8 but not 100% of the time
    - Tune eye's using serdes CTLE changes
    - Tune AOC output (doesn't normally help much)
  - ✓ Receiver errors gone but still not linking 100% of the time
    - Reduce Detect.Quiet wait time
- Successful link 100% of the time across all devices tested

Detect Quiet

- Reducing the detect.quiet time reduces optical module saturation
  - ✓ See Foil 12

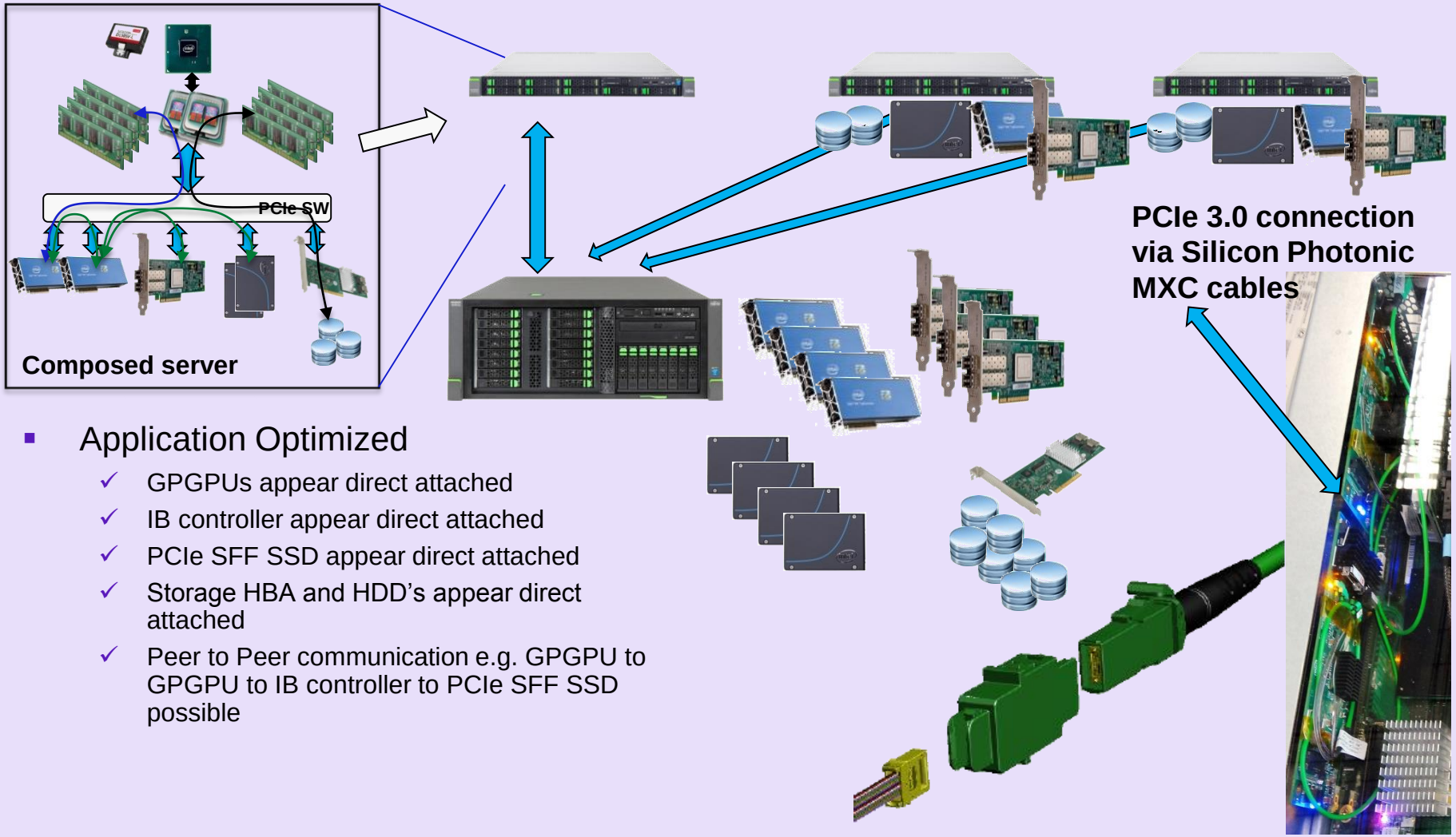




Which Optical Devices to Use

- We've tested many manufacturers optical devices including
  - ✓ Avago
  - ✓ Finisar
  - ✓ FCI
  - ✓ Samtec
- All have worked with appropriate serdes changes
  - ✓ Some only required minimal changes

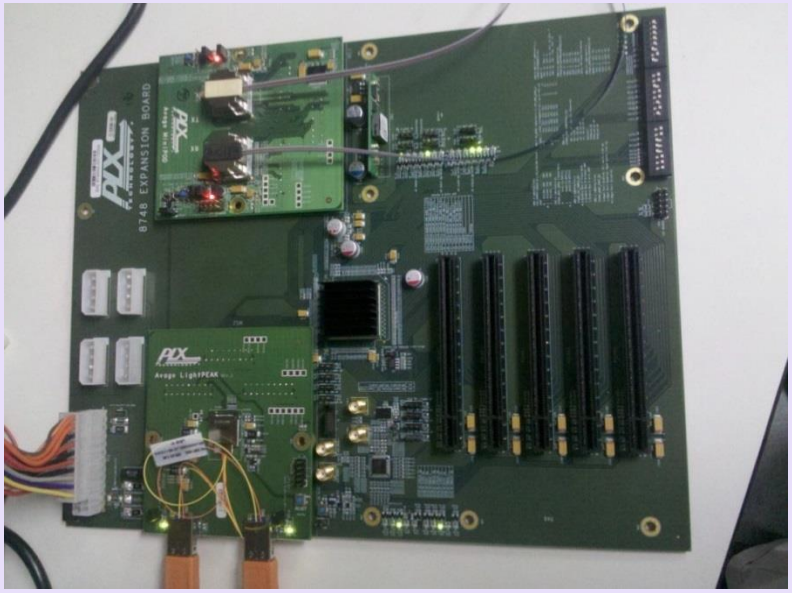
Flexible Application Optimized Server

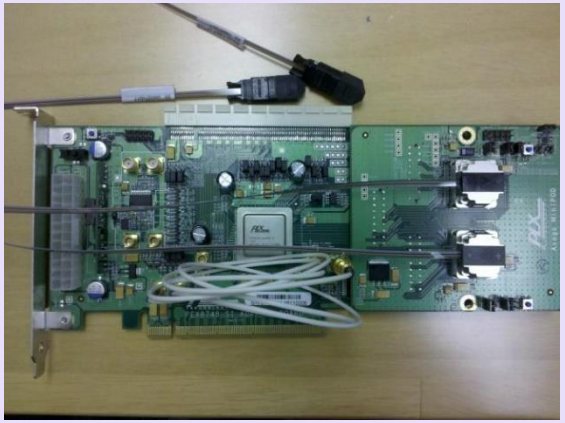


- Application Optimized
  - ✓ GPGPUs appear direct attached
  - ✓ IB controller appear direct attached
  - ✓ PCIe SFF SSD appear direct attached
  - ✓ Storage HBA and HDD's appear direct attached
  - ✓ Peer to Peer communication e.g. GPGPU to GPGPU to IB controller to PCIe SFF SSD possible

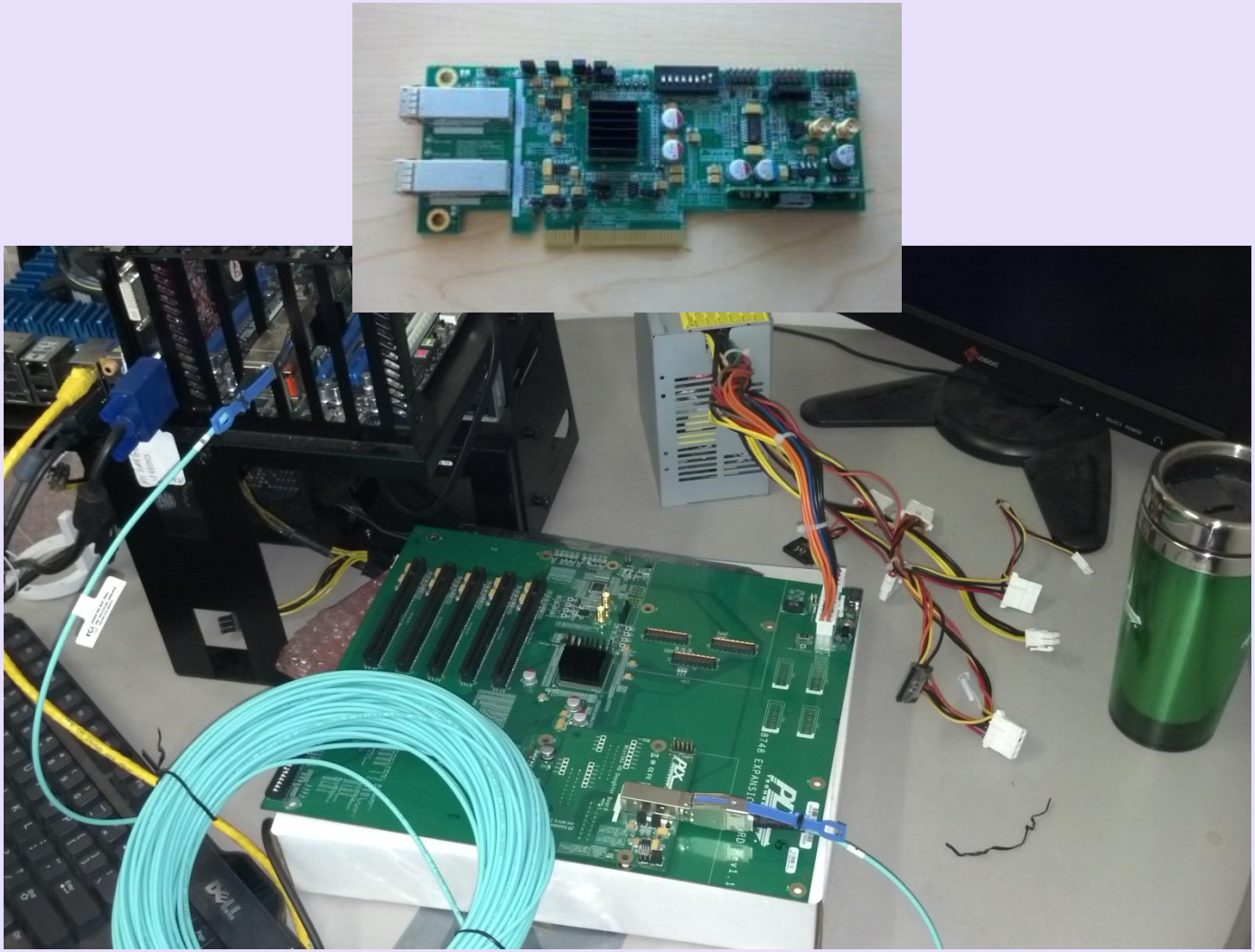
Examples of Test Boards

Test boards using Micropods

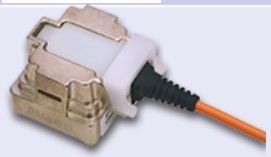
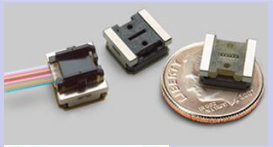




Test Board Using MiniSAS HD



Tested Optical Interfaces

Manufacturer	Optical/Connector Type	Part Number	Picture
Avago	QSFP+	AFBR-7QERxxZ	
Avago	MiniPOD	AFBR-81xxxZ/82xxxZ	
Avago	MicroPOD	AFBR-77DxxZ /78DxxZ	
Finisar	BOA	FBOPD10SL1L01	
Finisar	QSFP+	FTL410QxxC	

Tested Optical Interfaces

Manufacturer	Optical/Connector Type	Part Number	Picture
Finisar	C.Wire	FCBGD10CDxCxx	
FCI	MiniSAS HD	10117949-xxxxLF 10123196-xxxxxxx	
Samtec*	PCIe iPass	PCIEO-wGs-xxxx.x	

- Many other optical interfaces tested and shown to work by both Avago and various customers

*Some issues with 2.0 devices but this may have been an early firmware revision from Samtec. Customers have shown 3.0 to work with these cables.

Summary

- PCIe over optical cables works and has been used in production for many years now
- Currently serdes changes are required for the majority of optical cables
  - ✓ This requires some programmability in both PCIe link partners serdes
  - ✓ Testing and validation required for systems
  - ✓ Currently locks systems to known devices and optical interfaces
- New specs such as SRIS, OCuLink and the 3.0 Cable Spec should provide increased opportunities for linking over optical cables

Acknowledgements

- I'd like to thank the following for their assistance with this presentation:
  - ✓ Reginald Conley – Avago
  - ✓ Francesco Martini – Avago
  - ✓ Jim Ashbrook – Avago
  - ✓ Bernhard Schräder - Fujitsu

Thank you for attending the PCI-SIG Developers Conference

Israel 2015

For more information please go to www.pcisig.com

Appendix

- Useful Avago/PLX switch registers
- Receiver Detect and Electrical Idle Detect – 204h
- Downtrain disable
  - ✓ BE4h (station-based Phy Port Chicken Bits 0)
    - [5:0] (Downtrain disable – one bit per port of the station)
- Detect Quiet
  - ✓ 220h (station-based Physical Layer Function Control)
    - [9:8]=0 (reduce the Detect.Quiet Wait Time from 4 mS (default) to 0 mS)
- TCB disable
  - ✓ 224h (station-based PHY Layer Test 0)
    - [9] Ignore Compliance Receive TCB
  - ✓ 22Ch (station-based PHY Layer Chicken Bits)
    - [20] Ignore Hot Reset TCB
    - [21] Ignore Disable Link TCB
    - [22] Ignore Loopback TCB
    - [23] Ignore Disable Scrambling TCB
- Registers to monitor for errors
  - ✓ Link Status
    - 0x160[17] (VC0 Negotiation Pending) – most accurate method to determine link status
  - ✓ RX Detected Errors
    - 0x724 (Framing Errors)
    - 0xBC4[15:0] (Recovery Counter)
    - 0xBF0 (Receiver Errors)
    - 0xFAC (Bad TLP Counter)
    - 0xFB0 (Bad DLLP Counter)
  - ✓ TX Detected Errors
    - 0xFC4[12] (Replay Timer Timeout)
    - 0xFC4[8] (Replay Num Rollover)